NPRST

Navy Personnel Research, Studies, and Technology
5720 Integrity Drive • Millington, Tennessee 38055-1000 • www.nprst.navy.mil

research at work

NPRST-TN-11-1

November 2010

Practical Recommendations for Trait-level Estimation in NCAPS

Frederick L. Oswald, Ph.D.
Rice University

NPRST-TN-11-1
November 2010

Practical Recommendations for Trait-Level Estimation in the Navy Computer Adaptive Personality Scales (NCAPS)

Frederick L. Oswald, Ph.D.
Rice University

Reviewed, Approved, and Released by
David M. Cashbaugh
Director

Navy Personnel Research, Studies, and Technology
Bureau of Naval Personnel
5720 Integrity Dr.
Millington, TN 38055-1000
www.nprst.navy.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
30-11-2010	Technical Note	

4. TITLE AND SUBTITLE

Practical Recommendations for Trait-Level Estimation in the Navy Computer Adaptive Personality Scales (NCAPS)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Frederick P. Oswald, Ph.D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)

Rice University
6100 Main St., MS25
Houston, TX 77005

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1)
Bureau of Naval Personnel
5720 Integrity Drive
Millington, TN 38055-1000

10. SPONSOR/MONITOR'S ACRONYM(S)

NPRST/BUPERS-1

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

NPRST-TN-11-1

12. DISTRIBUTION / AVAILABILITY STATEMENT

A - Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report provides an analysis of Sailor Data obtained using the Navy Computer Adaptive Personality Scales (NCAPS), a tool intended to enhance selection and classification procedures by increasing commitment and performance, and reducing attrition. Analyses in the present report compared (a) existing complex item-scoring methods based on item response methods (IRT) to (b) two simpler scoring methods, one awarding points based on the expert rating of the endorsed item, the other giving one point for endorsing the item in the pair presented that reflects more of the given trait and zero points otherwise. Results indicated that simpler methods were relatively distinct from one another, and together, they predicted the vast majority of the variance in the more complex scoring method. Limitations and implications of these findings are discussed.

15. SUBJECT TERMS

NCAPS, IRT, computer adaptive, personality scales

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Genni Arledge
a. REPORT UNCLASSIFIED	**b. ABSTRACT** UNCLASSIFIED	**c. THIS PAGE** UNCLASSIFIED	UNLIMITED	29	**19b. TELEPHONE NUMBER** *(include area code)* 901-874-2115 (DSN 882)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Foreword

This report provides an analysis of Sailor data obtained using the Navy Computer Adaptive Personality Scales (NCAPS). The NCAPS measure is part of a suite of tools developed under the Whole Person Assessment program, a program geared toward modernizing Navy personnel selection and classification practices and algorithms, which at present rely almost exclusively on cognitive ability measures. In the test-centered world of personnel activities, there is the danger of developing more tests for their own sake, but the case for NCAPS could not be more clear: high levels of Sailor competence at the tip of the spear requires much more of Sailors than cognitive ability alone; it also requires high levels of vigilance, social orientation, dependability, self-reliance, and other personality traits reflected in NCAPS. Incorporating NCAPS into Navy selection and classification systems has the potential to increase Sailor satisfaction and performance, thereby increasing commitment and reducing attrition, two outcomes of critical interest to the US Navy.

Analyses and results from this report extend and qualify those from previous NCAPS efforts (e.g., Houston, Borman, Farmer, & Bearden, 2006) by examining the properties of NCAPS personality scales and its adaptive structure in more detail. Specifically, this report examines item exposure rates, the efficiency of item use based on IRT-based Expected A Posteriori (EAP) scoring, and a comparison of IRT-EAP scoring with much simpler scoring methods that may work just as well (stem-level scoring and dichotomous scoring). The cutting-edge nature of NCAPS testing will necessitate a series of efforts like this to push the technology further ahead.

This work was conducted under the support of Navy Personnel Research, Studies & Technology (NPRST, Millington TN). The author would like to acknowledge the work of Elizabeth M. Poposki, who assisted in the data analysis. The opinions contained herein are those of the author and do not reflect the official policy or position of the U.S. Navy, Department of Defense, or the U.S. Government. Please direct all correspondence to Fred Oswald, Department of Psychology, Rice University, 6100 Main St., MS25, Houston TX, 77005. E-mail foswald@rice.edu.

David M. Cashbaugh
Director

Contents

List of Tables

List of Figures

Introduction

This report describes the analysis of a large data set that contains Sailors' scores on ten personality traits measured by the Navy Computer Adaptive Personality Scales (NCAPS; see Table 1). As the name implies, NCAPS is a computerized personality test; it is adaptive in the sense that the computer tailors the presentation of test questions based on the Sailor's previous pattern of responses to test questions, and when fully operational, the system will stop the test whenever a Sailor's personality score on a trait, called an EAP (Expected A Posteriori) score, achieves a pre-specified standard of accuracy (i.e., psychometric reliability). The computerized format of NCAPS has the promise of thwarting Sailors' attempts to fake or look "too good" on the personality test, in two ways. First, NCAPS pairs stems together based on a Sailor's previous item responses, creating items (stem-pairs) that are tailored to each person, thereby enhancing test security by minimizing the exposure of item and stem content. In fact, it would be highly unlikely for any two NCAPS tests to be exactly the same. Second, NCAPS is delivered in a forced-choice format, which means that each personality item requires that the Sailor answer one of two options (stems)—meaning it is difficult to look "too good" for items where both stems are somewhat undesirable. Ultimately, the goal of NCAPS is to yield practical benefits to the Navy, first and foremost being the power of the measure to predict outcomes above and beyond existing personnel selection tools (e.g., ASVAB), yielding benefits of enhanced performance and reduced turnover. It is important to note that even when validity gains are modest, the benefits from those gains are multiplied across all Sailors in the selection system over time. Cost savings are also reaped from the NCAPS computer adaptive testing format by way of more efficient testing time, lower per-person testing costs, and rapid test updating.

Table 1
NCAPS Personality Traits

Adaptability/Flexibility (ADF)	191 stems
Attention to Detail (ADL)	164 stems
Achievement (AV)	108 stems
Dependability (DEP)	185 items
Dutifulness/Integrity (DUT)	152 stems
Social Orientation (SO)	114 stems
Self-Reliance (SRL)	199 items
Stress Tolerance (ST)	119 items
Vigilance (VIG)	106 items
Willingness to Learn (WTL)	156 stems

Note. Three-letter abbreviations in parentheses are used throughout the text and tables. 1,494 total item stems for these 10 NCAPS traits. Stems are paired together using IRT-based methods to form items.

Recent research by Houston et al. (2006) has analyzed NCAPS data, finding promise for NCAPS as a component of Navy personnel selection and classification systems—and more generally for the Whole Person Assessment efforts of the Navy. They reported that the 10 NCAPS personality scales were more reliable than an analogous personality measure that used a traditional (Likert-scale) scoring method. The two measures also showed comparably high validities for predicting an overall performance composite of peer ratings (r = .32 and r = .39 for NCAPS vs. traditional measurement, respectively). Additionally, these researchers found a similar level validity for NCAPS scores predicting supervisory ratings (r = .37). Perhaps their most interesting finding is how they found much *lower* validity for the traditional personality measure predicting the same supervisory ratings (r = .18). These broad findings provide initial evidence that NCAPS retains high levels of validity for supervisory ratings, perhaps because it is less prone to the sort of faking or impression management that encumbers traditionally scored personality tests. This hypothesis is consistent with higher validities that have been found both in lab and field studies that have examined other forced-choice personality test formats (e.g., Bartram, 2007; Jackson, Wroblewski, & Ashton, 2000). Future research might investigate more specific patterns of correlations between personality scales and dimensions of performance for peer ratings and for supervisor ratings to explain or qualify these results further.

NCAPS is not only adaptive and forced-choice in nature; it also invokes a complex scoring model based on item response theory (IRT). Typically, IRT models used in psychological testing assume a *dominance model*, meaning that people who possess higher levels of a trait should be more likely to answer items correctly (in the case of an ability test) or in a more positive direction (in the case of a personality trait). The dominance model is especially appropriate for cognitive ability testing, because generally speaking, a person who can answer hard items correctly on a cognitive ability test will tend to answer all easier items correctly as well; conversely, a person who cannot answer easy items correctly will be less likely to answer harder items correctly. In contrast with the dominance model, the IRT model used for NCAPS personality traits uses the *ideal-point* model (or unfolding model). The ideal-point model operationalizes the assumption that individuals will be more likely to endorse personality items that are "closest to" their actual trait level (see Coombs, 1950; Stark & Drasgow, 2002; Zinnes & Griggs, 1974). For instance, under the IRT ideal-point model, a person possessing an average level of Achievement is assumed to endorse more items that reflect an average level of the trait, and this person is less likely to endorse items reflecting a very high level of Achievement or a very low level of Achievement (note that the dominance model might instead suggest this person is more likely to endorse items reflecting low levels of achievement).

In short, as a Sailor responds to NCAPS items, the ideal-point model (a) estimates an individual's trait standing at that time and with that estimate (b) decides on the trait levels of the two stems that will get paired together to form the next forced-choice item that will achieve a more accurate measurement on the individual. It is this latter point that makes NCAPS an adaptive test: Measurement stops when the estimated accuracy of an individual's trait level has reached a pre-specified level of accuracy, or when the pre-established maximum number of item pairs per trait is reached, whichever comes first. When the large sample size requirements are met to yield accurate and interpretable

IRT statistics, then ideal-point IRT models tend to show better empirical fit to personality test data over traditional IRT models. But the increase in fit may be in part because there are more parameters to estimate, which increases the complexity of the model. There is always a tradeoff between model complexity and model fit: more complex models tend to achieve better fit, but they also have a greater chance of capitalizing on idiosyncrasies of the particular set of data being modeled. Therefore, model fit is not the single goal of a good model; a good model is also appropriately parsimonious (Pitt & Myung, 2002; Preacher, 2006). Finding an appropriate tradeoff between model fit and model complexity can be tricky. Although more complex models may be theoretically appealing than simpler models and also fit a single sample of data better, the parameter estimates from simpler models have a better chance of fitting the data from new samples and testing occasions; simpler models also adhere to the principle of parsimony in science (i.e., Occam's Razor: adopt the model that is the most parsimonious yet accounts for the data reflecting the phenomenon of interest). Recent work on IRT ideal-point models has argued for more research examining the tradeoff between model fit and model complexity (Ferrando, 2006).

Regarding NCAPS in particular, if the complexity of the IRT ideal-point model does not purchase a commensurate increase in psychometric reliability or criterion-related validity, then the complexity of this model can justifiably be "shaved down" with Occam's razor to arrive at a simpler model where NCAPS items are scored by a simpler method. The general literature on IRT suggests that trait-level estimation of individuals using the simpler classical test theory (CTT) approach (e.g., operationalizing a trait score as the average response across relevant items) often correlates with its more complex IRT counterpart as high as .97 and even higher (see Fan, 1998, in the context of ability testing based on dominance IRT models). In their popular IRT text, Embretson and Reise (2000, p. 324) raise this same point within the context of measuring personality and attitudes:

> Our observations are that raw scores and trait level estimates always correlated greater than .95, and no one has ever shown that in real data a single psychological finding would be different if IRT scores were used rather than raw scale scores. Without such demonstrations, arguments of IRT supporters will fall on deaf ears.

Although there is no gold standard for how to make an optimal tradeoff between model complexity and model fit, it is informative to determine whether the additional model complexity of the IRT ideal point model embedded within the NCAPS technology might translate into additional practical benefit in a Navy personnel testing system, or whether simpler models might be preferred instead, because they capture the phenomenon of interest just as well as complex models while retaining all benefits. This paper makes one specific attempt at evaluating the complexity-fit tradeoff by determining whether a scoring model that is much simpler the IRT ideal-point model might provide scores that are similar to their IRT counterparts (i.e., are highly correlated) and therefore potentially just as useful in terms of their predictive validity.

Methods

Data Set

The data set under investigation comprises NCAPS data for 8,956 U.S. Navy Sailors.[1] All participants in this sample completed at least 7 items across all 10 traits; at least 96 percent of participants had complete data on each trait; and 83 percent answered all 12 items across all 10 traits.

For each Sailor there are data on the following

- The ID corresponding to the stems that were paired together to create each item presented.
- The response selected for each item. Each Sailor responded to 12 items for 10 traits, or 120 items total. The number of administered items is a large fixed number in this developmental phase. The NCAPS system is still adaptive in this case, because the system still selects item-stem pairs based on previous EAP estimates, but it will become more adaptive once operational, by shortening the test based on an acceptable level of error (posterior standard deviation [PSD], see below).
- The estimated personality trait score or EAP (Expected A Posteriori) score on the trait to which the item belongs (i.e., the EAP is re-estimated after each item response for a given trait).
- The estimate of the standard error (posterior standard deviation, or PSD) for each Sailor's EAP.
- The final NCAPS EAPs and PSDs for each Sailor across each of 10 traits (see Table 1 for the list of traits).

Item stem content was confidential and not available to the author; therefore it was not considered here, though examination of this content associated with the current analyses may certainly inform future efforts to understand and further refine the NCAPS measure.

[1] These data comprise a subset of the full data base. Cases were deleted as follows (per Hubert Chen, personal communication): Deleted the first 44 test cases; delete participants prior to 11/10/2005; delete incomplete data; deleted cases where the total survey administration time was too fast (< 15 minutes). In addition to these deletions, I deleted N = 1,425 cases (about 15%) that did not complete at least 7 NCAPS items for each of the 10 traits. This ensures that all final EAPs in the data have relatively high reliability, and it simultaneously avoids the problem of attempting to impute missing data on the basis of complex IRT-based item selection and scoring. It is doubtful that including these cases materially affects the results presented, though a reanalysis involving these cases with missing data is welcome.

Item Stem Exposure

The NCAPS data base that was analyzed contains 1,494 item stems distributed across the 10 measured personality traits, where within each trait, the NCAPS computer algorithm selects 2 stems and pairs them to form an item. Given a trait with p item stems, there are as many as $p(p\text{-}1)/2$ unique pairs of stems that can form an item. Adaptability/Flexibility has 191 stems, for example, and therefore 18,145 items (stem pairs) are mathematically possible. However, there are constraints placed on how stems are selected that reduce the number of possible pairs seen in practice by a fair amount. Specifically, item stems are selected based on the person's current EAP (IRT-based trait score) for a given personality trait, such that a response to the selected item will provide as much additional diagnostic information as possible when re-estimating the EAP. If both of the selected stems were to measure a level of the personality trait that is much higher or much lower than the person's current EAP estimate, then generally the item response would be much less informative than a response to an item whose stems measure levels that are either close to or straddle the person's current EAP.

The choice of item stems in NCAPS is non-random and can lead to differential rates of item stem use. A prime example that applies to every test-taker is the first NCAPS item administered for a given trait, when a person does not yet have an EAP value: given that no previous items have been administered, there is no prior EAP information, and every person's initial EAP is estimated at the overall mean. Therefore, the rule for selecting and pairing item stems to generate the first item for each NCAPS trait should be the same for every test taker. This suggests that during the initial presentation of NCAPS items, some item stems will be administered (or "exposed") more often than other item stems. There may be other reasons for some stems to receive more exposure than others. For instance, because personality traits are approximately normally distributed, fewer people have EAPs at the extremes; therefore, items providing more precise measurement only at the extremes would be exposed less). Another source of undue item exposure comes from any test-takers having pre-existing knowledge of the item stems which provide desirable test responses. To the extent individuals have prior knowledge or know how to "fake" the test, this has the potential to compromise the predictive validity of the test. Based on the aforementioned potential reasons for differential item exposure and need for test security, stem exposure was investigated within the current data set.

Figure 1 (at the end of this report) displays panels of histograms across the 10 NCAPS traits in the data base. For each trait, the x-axis represents the percentage of time that a stem was exposed across all administrations of NCAPS, and the y-axis represents the percentage of stems with a level of item exposure on x. One can see that most stems for a trait (between 70–80%) are exposed only 1 percent of the time. Some stems are exposed 3–4 percent of the time; generally, this happens more often for traits with fewer stems (e.g., Achievement, Vigilance) than for traits with more stems (e.g., Adaptability, Self-Reliance).

Results and Discussion

Relationship between Initial IRT-EAP Scores and Final IRT-EAP Scores

Recent research in IRT-CAT has indicated that while trait estimation and item administration should be dependent on previous test-taker responses, an IRT-CAT program can also be overly sensitive to previous item responses. In other words, if test-takers do not respond initially in ways consistent with their underlying trait score, they may not have enough test-taking opportunity during the remainder of the test to reveal their true underlying trait score (Chang & Ying, 2007).

Findings from NCAPS in Table 2 show that the EAP estimated from just the first few NCAPS items that are administered correlated substantially with the final EAP estimated after 12 items (i.e., from .71 to .82 after 3 items); however, the correlation is still low enough to allow NCAPS to adjust each person's EAP after each item response. Let us investigate how much benefit might be gained by responding to each additional item for a given trait. For a given trait, we will consider the correlation between (a) the EAP estimated after responding to each item on a trait and (b) the final EAP that is estimated after taking all items for that trait. A reliability value of .70 (a typical benchmark for internal consistency reliability under classical test theory) corresponds to a correlation of .84—the square root of .70. Table 2 shows that using this benchmark, an EAP score based on 5 items would be a sufficient indicator of the final EAP score based on all 12 items. For even a more stringent level of reliability of .85, the corresponding correlation is .92, and Table 2 suggests that the marginal gains of additional items for estimating the EAP score would be very slight after 7 items. By these standards, it appears that using 8 items per trait would generally achieve sufficient reliability for most practical purposes.

Table 2
Correlation between Initial IRT-EAP Scores and Final IRT-EAP Scores

EAP	ADF	ADL	AV	DEP	DUT	SO	SRL	ST	VIG	WTL
1 item	.55	.55	.47	.57	.49	.49	.54	.58	.54	.51
2 items	.68	.71	.61	.72	.64	.65	.67	.73	.69	.68
3 items	.77	.80	.71	.81	.72	.76	.77	.82	.78	.76
4 items	.83	.85	.79	.86	.77	.82	.83	.87	.84	.83
5 items	.88	.89	.84	.90	.83	.87	.88	.91	.89	.87
6 items	.91	.92	.88	.93	.87	.90	.91	.94	.92	.90
7 items	.93	.94	.92	.95	.91	.93	.93	.96	.94	.93
8 items	.95	.96	.94	.97	.94	.95	.95	.97	.96	.95
9 items	.97	.97	.96	.98	.96	.97	.97	.98	.97	.97
10 items	.98	.98	.97	.99	.97	.98	.98	.99	.99	.98
11 items	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99

Note. N = 8,956. Some participants did not respond to all items (see text). Those who did not respond to an item received the same EAP that they were previously assigned.

Note that the analyses summarized in Table 2 provide some CTT-like evidence regarding the reliability of IRT-based EAP trait scores (i.e., unconditional reliability, in contrast with each person's PSD). One purported advantage of IRT over CTT is that reliability estimates are conditional on a person's underlying trait score; these conditional reliability estimates are called *posterior standard deviations* (PSD) that are estimated alongside each person's IRT-based EAP score. Although similar reliability estimates can be computed as part of CTT (e.g., an alpha reliability conditioned on a person's total score), it is not a featured element of the CTT model; psychometric approaches using CTT typically provide an overall (unconditional) estimate of reliability. A more straightforward comparison between EAP scoring (based on IRT) and simpler scoring methods has been be examined here, but future work might find benefit in comparing IRT and CTT models of NCAPS more closely with each other.

The conclusions here must also be qualified in another important way: the NCAPS data base for the present analyses lacks criterion data. Criterion-related validities for EAP scores at different test lengths might suggest different cutoffs than reliability analyses do. Specifically, there may be a larger or smaller number of NCAPS items suggested for each trait, beyond which additional items add little to criterion-related validity. Previously-mentioned research had validity data and conducted these types of analyses (Houston et al., 2006, Chapter 5); those analyses could be usefully extended in future research.

IRT-EAP Scoring vs. Simpler Scoring Procedures

In addition to the previous reliability analysis, a key purpose for the current re-analysis of NCAPS data was to determine whether scoring items in a simpler manner might yield similar personality trait scores as the current complex IRT-based scoring system.

First, a piece of background information is required: Each item stem in NCAPS was rated by Navy subject matter experts (SMEs), where ratings indicate how high (or how low) a stem is on a given personality trait.[2] The SME ratings were averaged for each stem, and IRT method uses the average SME ratings for each stem to determine which ones should be paired to form an item, given a test-taker's current estimated EAP score on a trait.

Two alternative methods of scoring were investigated; they are much simpler than IRT-based EAP scoring. The first method of scoring is called the *stem-level scoring* method: For each of the 10 NCAPS traits, there are mean SME ratings associated with the stems that a Sailor selected for all 12 items; this method takes the mean of those 12 ratings. One could add up the mean SME ratings across the 12 endorsed stems instead; however, averaging is better, because some Sailors have missing data, and with averaging instead of adding, lower scores are not merely due to missing data. For example, the current data set contains Sailors' responses to 12 NCAPS items reflecting the personality trait of Self Reliance. Across all Self Reliance items, Sailors who

[2] Note that the "level" that the item stems reflect a personality trait can also be determined empirically, such as through IRT methods (i.e., in typical IRT applications, the method will "score" or scale items as well as people on the construct of interest).

consistently endorse the stem reflecting a higher trait level (i.e., higher SME rating) will receive the highest level of Self Reliance score that can be estimated. At the other extreme are Sailors who always choose stems with the lower level of self-reliance; these Sailors would have the lowest score possible. Of course, most Sailors will select the higher-level stem sometimes, and other times they will select the lower-level stem, so their scores will be situated between the two aforementioned extremes.

The second method of scoring is called the *dichotomous scoring* method, awarding a Sailor 1 point for endorsing the higher-level stem in a pair and 0 points for the lower-level stem, then averaging across the number of items responded to for the given trait. Taking the previous example, Sailors who always select the stem with the higher level of self reliance would receive a score of 1.0. Most scores will fall between 0.0 and 1.0.

Generally speaking, item information tends to increase in direct proportion to the distance between the stems (Stark & Drasgow, 2002). If this principle is taken to its extreme, then the most useful stems tend to be those with stems at the highest and lowest ends of the personality trait continuum. Most mean SME ratings of NCAPS stems, in fact, fall at the upper and lower ends of each continuum of the 12 traits. If the usefulness of stems at these extremes overwhelms the usefulness of adapting items (stem pairs) to a person's given EAP, then both stem-scoring and dichotomous scoring serve as viable and simpler alternatives to EAP scoring. Conversely, if item stems need to be tailored to an EAP, then these simpler scoring methods will offer less benefit.

Tables 3, 4, and 5 provide correlations between IRT-based EAP scoring, stem-level scoring, and dichotomous scoring methods, revealing some interesting patterns of relationship. Table 3 shows that EAP scoring and stem-level scoring demonstrate non-trivial levels of convergence on their corresponding traits,[3] with correlations ranging between .51 and .69. At first glance, these correlations may not seem that interesting: a given NCAPS item (stem-pair) is based on a Sailor's current EAP score; therefore, stem-level scores might correspond with EAP scores even for random responses. We know that this cannot be entirely true because of results for dichotomous scoring that, together with these results, indicate that both the level and direction of the item stem choices matter.

[3] Positive correlations for non-corresponding traits simply reflect the likely fact that all the traits are positively correlated and any reliable measures of the dimensions will reflect that.

Table 3
Correlation between 10 NCAPS Traits:
IRT-EAP Scores (Rows) vs. Stem-level Scores (Columns)

	ADF	ADL	AV	DEP	DUT	SO	SRL	ST	VIG	WTL
ADF	**.63**									
ADL	.19	**.68**								
AV	.21	.26	**.52**							
DEP	.20	.36	.21	**.66**						
DUT	.15	.23	.15	.24	**.55**					
SO	.19	.15	.12	.13	.09	**.59**				
SRL	.08	.04	.09	.03	.01	-.06	**.61**			
ST	.24	.20	.19	.20	.11	.12	.09	**.69**		
VIG	.20	.30	.20	.28	.16	.11	.08	.19	**.51**	
WTL	.21	.24	.18	.23	.15	.14	.00	.19	.18	**.63**

Note. *N* = 8,956. ADF = Adaptability/Flexibility, ADL = Attention to Detail, AV = Achievement, DEP = Dependability, DUT = Dutifulness/Integrity, SO = Social Orientation, SRL = Self-Reliance, ST = Stress Tolerance, VIG = Vigilance, WTL = Willingness to Learn. Correlations with an absolute value > .01 are statistically significant ($p < .01$).

Table 4 summarizes the correlations between IRT-based EAP scoring with dichotomous scoring. Here, the convergent correlations between corresponding traits are between .77 and .88—high in absolute magnitude and higher than the correlations between EAP scoring with stem-level scoring. The correlations might be high for the reason previously cited, that generally, responses to multiple items with stem levels that are far apart tend to provide more information about a Sailor's personality trait level than stems than items with stems that are closer together.

Table 4
Correlation between 10 NCAPS Traits:
IRT-EAP Scores (Rows) vs. Dichotomous Scores (Columns)

	ADF	ADL	AV	DEP	DUT	SO	SRL	ST	VIG	WTL
ADF	**.81**									
ADL	.30	**.83**								
AV	.35	.39	**.77**							
DEP	.30	.51	.37	**.87**						
DUT	.24	.34	.27	.38	**.77**					
SO	.32	.19	.20	.21	.21	**.82**				
SRL	.11	.07	.17	.09	.00	-.12	**.80**			
ST	.40	.29	.31	.35	.21	.24	.14	**.87**		
VIG	.32	.45	.37	.48	.31	.22	.14	.36	**.88**	
WTL	.35	.34	.32	.38	.30	.26	.02	.34	.35	**.82**

Note. N = 8, 956. ADF = Adaptability/Flexibility, ADL = Attention to Detail, AV = Achievement, DEP = Dependability, DUT = Dutifulness/Integrity, SO = Social Orientation, SRL = Self-Reliance, ST = Stress Tolerance, VIG = Vigilance, WTL = Willingness to Learn. Convergent correlations are in boldface on the main diagonal. Correlations with an absolute value > .01 are statistically significant (p < .01).

Findings for IRT-based EAP scoring converging with both stem-level scoring and dichotomous scoring are limited when each source of convergence is examined independently. It is when they are taken together that the findings for each are compelling, because one learns that the simple scoring methods are not redundant. Table 5 shows low correlations between scoring methods, with correlations on corresponding traits ranging from -.07 to .26. These clearly are different methods, then, for scoring NCAPS, yet they both correlate highly with EAP scoring. This suggests that both simple methods, taken *together*, may explain the usefulness of EAP scoring: the EAP scoring procedure accounts for the rated *level* of the item stems (as in stem-level scoring), but also its discernible *direction* or *valence* (as in dichotomous scoring), due to the fact that the stems are presented to the test-taker in pairs. Table 6 indicates that both contribute uniquely to EAP scoring, yet together account for almost all (95–99%) of the variance in EAP scoring. Additional analyses might investigate the spacing or distance between the levels of the stems that are paired together to determine how strongly stem distances for each item is related to stem-level scoring, dichotomous scoring, and the accuracy of the Sailor's EAP (e.g., as measured through the PSD or through other simpler methods).

Table 5
Correlation between 10 NCAPS Traits: Stem-level Scores (Rows) vs. Dichotomous Scores (Columns)

	ADF	ADL	AV	DEP	DUT	SO	SRL	ST	VIG	WTL
ADF	**.08**									
ADL	.17	**.18**								
AV	.14	.15	**-.12**							
DEP	.15	.26	.17	**.23**						
DUT	.08	.14	.08	.16	**-.07**					
SO	.14	.07	.08	.08	.09	**.05**				
SRL	.07	.03	.10	.04	.00	-.05	**.04**			
ST	.19	.14	.13	.18	.10	.13	.06	**.26**		
VIG	.11	.15	.11	.18	.09	.08	.04	.13	**.06**	
WTL	.16	.17	.13	.19	.13	.13	-.02	.16	.15	**.10**

Note. N = 8, 956. ADF = Adaptability/Flexibility, ADL = Attention to Detail, AV = Achievement, DEP = Dependability, DUT = Dutifulness/Integrity, SO = Social Orientation, SRL = Self-Reliance, ST = Stress Tolerance, VIG = Vigilance, WTL = Willingness to Learn. Convergent correlations are in boldface on the main diagonal. Correlations with an absolute value > .01 are statistically significant (p < .01).

Table 6
Stem-level Scores and Dichotomous Scores Predicting Variance in IRT-EAP Scores

	$r_{\text{dichot-level}}$ (Table 5)	$r_{\text{EAP-level}}$ (Table 3)	$r_{\text{EAP-dichot}}$ (Table 4)	R^2_{level}	R^2_{dichot}	R^2_{both}
ADF	.08	.63	.81	.40	.66	.98
ADL	.18	.68	.83	.46	.69	.98
AV	-.12	.52	.77	.27	.59	.97
DEP	.23	.66	.87	.44	.76	.98
DUT	-.07	.55	.77	.30	.59	.96
SO	.05	.59	.82	.35	.67	.97
SRL	.04	.61	.80	.37	.64	.97
ST	.26	.69	.87	.48	.76	.99
VIG	.06	.51	.88	.26	.77	.98
WTL	.10	.63	.82	.40	.67	.98

Note. N = 8, 956. ADF = Adaptability/Flexibility, ADL = Attention to Detail, AV = Achievement, DEP = Dependability, DUT = Dutifulness/Integrity, SO = Social Orientation, SRL = Self-Reliance, ST = Stress Tolerance, VIG = Vigilance, WTL = Willingness to Learn; dichot = dichotomous scores, level=stem-level scores, EAP=IRT-EAP scores. All correlations and R^2 values are statistically significant (p < .01).

Final Considerations

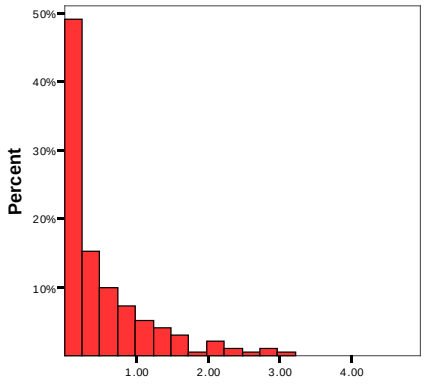
This report concludes with two caveats. The first caveat of comparing EAP scoring to a simpler method of scoring bears repeating in this brief report: NCAPS not only implements the IRT ideal-point model to score the item responses; it also implements IRT to determine the nature and ordering of the item pairs. In other words, when a Sailor takes NCAPS, his/her responses on all previous items for a given NCAPS trait are scored via IRT procedures and helps the IRT model determine the type of item the person will receive next. The virtue of this procedure for adaptive testing is also a caveat when comparing IRT to other methods using adaptive test data: The comparability between methods is affected by the fact the IRT method itself was used to select which items were administered. Future research that explores whether the complexity of NCAPS (and IRT modeling) is necessary in order to be of practical use should compare the present NCAPS administration method to more traditional or non-adaptive methods where each person either receives the same set of NCAPS items, or receives items independent of their EAPs (e.g., only item pairs with stems with levels at opposite extremes). The present NCAPS findings are quite compelling despite this caveat and need for additional research. It appears that in the context of the NCAPS personality test, simpler scoring methods appear to be just as useful as highly complex IRT-based methods.

The second caveat is that incorporating performance data or other criterion data relevant ot personality would inform the validity of NCAPS. When a test like NCAPS is in operational use, it is useful to understand how different test-scoring methods affect criterion-related validity. It would be useful to know whether simpler scoring methods yield scores with criterion-related validities that are comparable to complex IRT-based methods. It is possible for two methods with highly correlated results to provide drastically different validities (Wang & Stanley, 1970). However, there is no theoretical reason for expecting such differences, and practically speaking any differences are likely to be limited by the nature of the criterion data, not by the scoring method itself.

Figure 1. Exposure of NCAPS stems across traits.

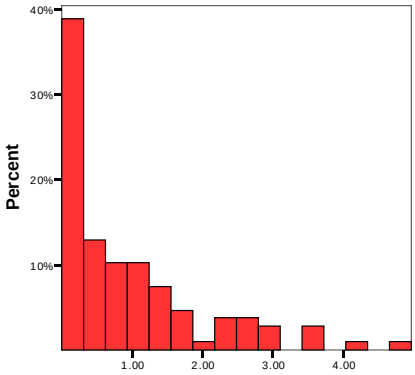
The graphs below provide (a) Stem exposure rates on the *x*-axis and (b) percentage of stems with that exposure rate on the *y*-axis. Across facets, most stems are exposed less than .25 percent of the time, but some stems are exposed up to 4 percent of the time.

Adaptability/Flexibility (191 stems)



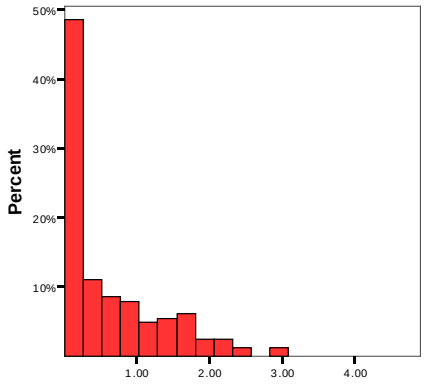
Item exposure over total item exposures by domain for ADF

Achievement (108 stems)



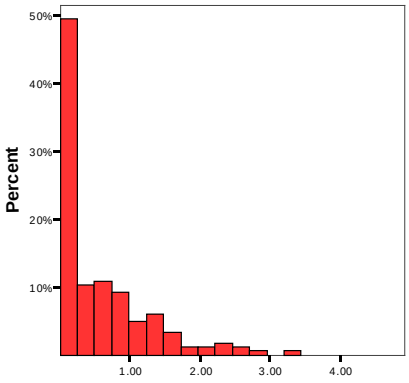
Item exposure over total item exposures by domain for AV

Attention to Detail (164 stems)



Item exposure over total item exposures by domain for ADL

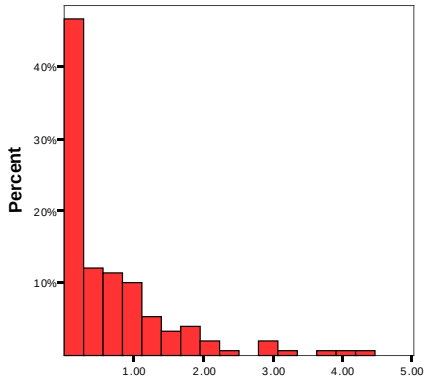
Dependability (185 stems)



Item exposure over total item exposures by domain for DEP

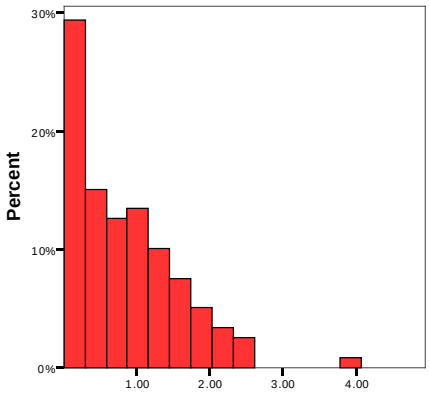
Figure 1. Exposure of NCAPS stems across traits (continued).

Dutifulness/Integrity (152 stems)



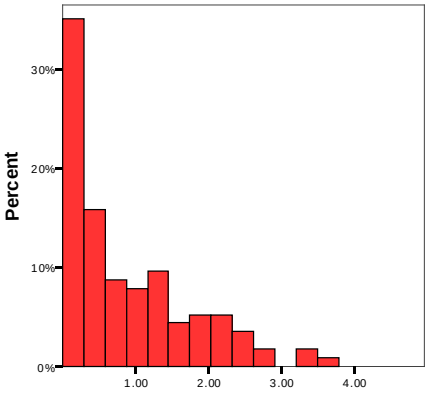
Item exposure over total item exposures by domain for DUT

Stress Tolerance (119 stems)



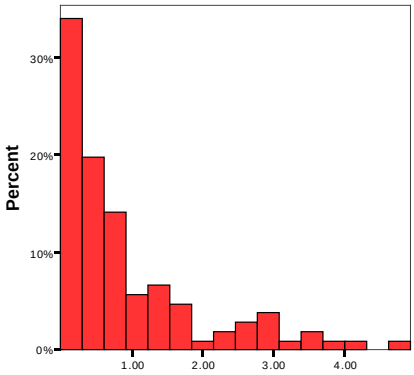
Item exposure over total item exposures by domain for ST

Social Orientation (114 stems)

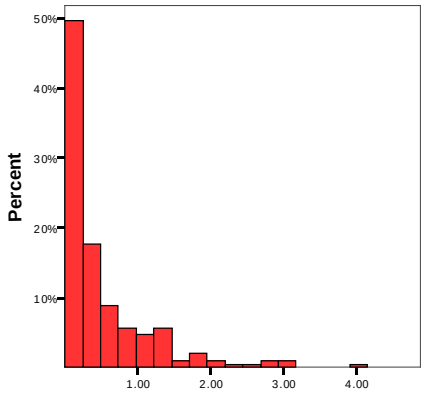
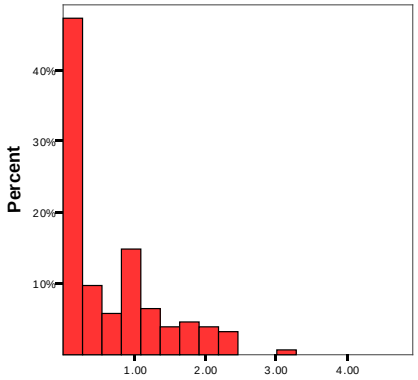


Item exposure over total item exposures by domain for SO

Vigilance (106 stems)



Item exposure over total item exposures by domain for VIG

Self-Reliance (199 stems)	*Willingness to Learn (156 stems)*
	
Item exposure over total item exposures by domain for SRL	**Item exposure over total item exposures by domain for WTL**

References

Bartram, D. (2007). Increasing validity with forced-choice measurement formats. *International Journal of Selection and Testing, 15,* 263-272.

Chang, H-H., & Ying, Z. (2007). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73,* 441-450.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review,* 57, 145-158.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychology Measurement, 58,* 357-381.

Ferrando, P. J. (2006). Two item response theory models for analyzing normative forced-choice personality items. *British Journal of Mathematical and Statistical Psychology, 59,* 379-395.

Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden, R. M. (2006). Development of the Navy Computer Adaptive Personality Scales (NCAPS). Millington, TN: Navy Personnel Research, Studies, and Technology.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13,* 371-388.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6,* 421-425.

Preacher, K. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41,* 227-259.

Stark, S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26,* 208-227.

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika, 39,* 327-350.

Distribution

AIR UNIVERSITY LIBRARY
ARMY RESEARCH INSTITUTE LIBRARY
ARMY WAR COLLEGE LIBRARY
CENTER FOR NAVAL ANALYSES LIBRARY
HUMAN RESOURCES DIRECTORATE TECHNICAL LIBRARY
JOINT FORCES STAFF COLLEGE LIBRARY
MARINE CORPS UNIVERSITY LIBRARIES
NATIONAL DEFENSE UNIVERSITY LIBRARY
NAVAL HEALTH RESEARCH CENTER WILKINS BIOMEDICAL LIBRARY
NAVAL POSTGRADUATE SCHOOL DUDLEY KNOX LIBRARY
NAVAL RESEARCH LABORATORY RUTH HOOKER RESEARCH LIBRARY
NAVAL WAR COLLEGE LIBRARY
NAVY PERSONNEL RESEARCH, STUDIES, AND TECHNOLOGY SPISHOCK LIBRARY (3)
OFFICE OF NAVAL RESEARCH (CODE 34)
PENTAGON LIBRARY
USAF ACADEMY LIBRARY
US COAST GUARD ACADEMY LIBRARY
US MERCHANT MARINE ACADEMY BLAND LIBRARY
US MILITARY ACADEMY AT WEST POINT LIBRARY
US NAVAL ACADEMY NIMITZ LIBRARY